Progetto di Linguistica Computazionale

A.A. 2018/2019

Linee guida

Obiettivo:
Realizzazione di due programmi scritti in Python che utilizzino i moduli presenti in Natural Language Toolkit per leggere due file di testo in inglese, annotarli linguisticamente, confrontarli sulla base degli indici statistici richiesti ed estrarne le informazioni richieste.

Fasi realizzative:
Create due corpora in inglese, di almeno 5000 token ciascuno, contenenti testi estratti rispettivamente da commenti positivi e negativi di prodotti o servizi venduti su siti on-line. Esempi di siti sono: Amazon (www.amazon.co.uk), Booking (www.booking.com), TripAdvisor (www.tripadvisor.co.uk). I commenti possono essere distinti tra positivi e negativi sulla base delle meta informazioni che ogni sito predispone vicino ai commenti (*stelle* per Amazon, *più* e *meno* per Booking, etc..). I corpora devono essere salvati in due file di testo semplice in codifica utf-8.
Sviluppate due programmi che prendono in input i due file da riga di comando, che li analizzano linguisticamente fino al Part-of-Speech tagging e che eseguono le operazioni richieste.

Programma 1 - Confrontate i due testi sulla base delle seguenti informazioni statistiche:

- il numero totale di frasi e di token;
- la lunghezza media delle frasi in termini di token e la lunghezza media delle parole in termini di caratteri;
- la grandezza del vocabolario e la Type Token Ratio (TTR) all'aumentare del corpus per porzioni incrementali di 1000 token (1000 token, 2000 token, 3000 token, etc.);
- la grandezza delle classi di frequenza $|V_3|$, $|V_6|$ e $|V_9|$ sui primi 5000 token;
- il numero medio di Sostantivi, Aggettivi e Verbi per frase.
- la *densità lessicale*, calcolata come il rapporto tra il numero totale di occorrenze nel testo di Sostantivi, Verbi, Avverbi, Aggettivi e il numero totale di parole nel testo (ad esclusione dei segni di punteggiatura marcati con POS "," "."):

  $(|Sostantivi|+|Verbi|+|Avverbi|+|Aggettivi|)/(TOT-(|.|+|,|))$.

Programma 2 - Per ognuno dei due corpora estraete le seguenti informazioni:

- estraete ed ordinate in ordine di frequenza decrescente, indicando anche la relativa frequenza:
  - i 20 token più frequenti escludendo la punteggiatura;
  - i 20 Sostantivi più frequenti;
  - i 20 Aggettivi più frequenti;
  - i 20 bigrammi di token più frequenti che non contengono punteggiatura, articoli e congiunzioni;
  - le 10 PoS (Part-of-Speech) più frequenti;
  - i 10 bigrammi di PoS (Part-of-Speech) più frequenti;
  - i 10 trigrammi di PoS (Part-of-Speech) più frequenti;

- estraete ed ordinate in ordine decrescente i 20 bigrammi composti da Aggettivo e Sostantivo (dove ogni token deve avere una frequenza maggiore di 2):
  - con frequenza massima, indicando anche la frequenza di ogni parola che compone il bigramma;
  - con *probabilità congiunta* massima, indicando anche la relativa probabilità;
  - con *forza associativa* massima (calcolata in termini di *Local Mutual Information*), indicando anche il relativo valore;

- le due frasi con probabilità più alta, dove la probabilità della prima frase deve essere calcolata attraverso un modello di Markov di ordine 0 mentre la seconda con un modello di Markov di ordine 1. I due modelli devono usare le distribuzioni di frequenza estratte dal corpus che contiene le frasi, le frasi devono essere lunghe minimo 6 e massimo 8 token e ogni token deve avere una frequenza maggiore di 2;

Risultati del progetto:
perché il progetto sia giudicato idoneo, devono essere consegnati:

a. i due file di testo contenenti i corpora;
b. i programmi ben commentati scritti in Python;
c. i file di testo contenenti l'output dei programmi.

Date di consegna del progetto: il progetto deve essere consegnato per posta elettronica a felice.dellorletta@ilc.cnr.it e alessandro.lenci@unipi.it almeno una settimana prima dello scritto di ogni appello per poter essere considerato valido per l'appello.
NB: il progetto **DEVE** essere svolto individualmente.